



# Water Segmentation via Asymmetric Multiscale Interaction Network

Jianzhuo Chen<sup>1</sup>, Tao Lu<sup>1(✉)</sup>, Yanduo Zhang<sup>1,3</sup>, Wenhua Fang<sup>1</sup>, Xiya Rao<sup>1</sup>,  
and Mingming Zhao<sup>2</sup>

<sup>1</sup> Wuhan Institute of Technology, Wuhan, China  
lut@wit.edu.cn

<sup>2</sup> Wuhan Fiberhome Technical Services Co., Ltd., Wuhan, China

<sup>3</sup> Hubei Three Gorges Laboratory, Yichang, China

**Abstract.** It is important to observe and split water region to help acquire the water quality and supervise water environment. Water segmentation is a task to separate water region from images. Due to the specular nature of the water surface, various types of reflections usually appear on the water surface, which can change significantly with weather and lighting changes, it is difficult for general segmentation to work. According to the characteristics of waters, i.e. wide area and reflection, we propose a asymmetric interaction module (AIM) converge the features to a larger receptive field. Further, with this powerful module, we design the asymmetric multiscale interaction network, which can maintain the features of each scale and reassign the weights of features at different scales. We conduct extensive experiments on Hubei water dataset we constructed, The results show the framework effectively improves the accuracy of water segmentation and greatly improves the visual effect of segmentation, which is 5.9% higher in self-made dataset with advanced methods.

**Keywords:** Water segmentation · Hubei water dataset · Asymmetric interaction

## 1 Introduction

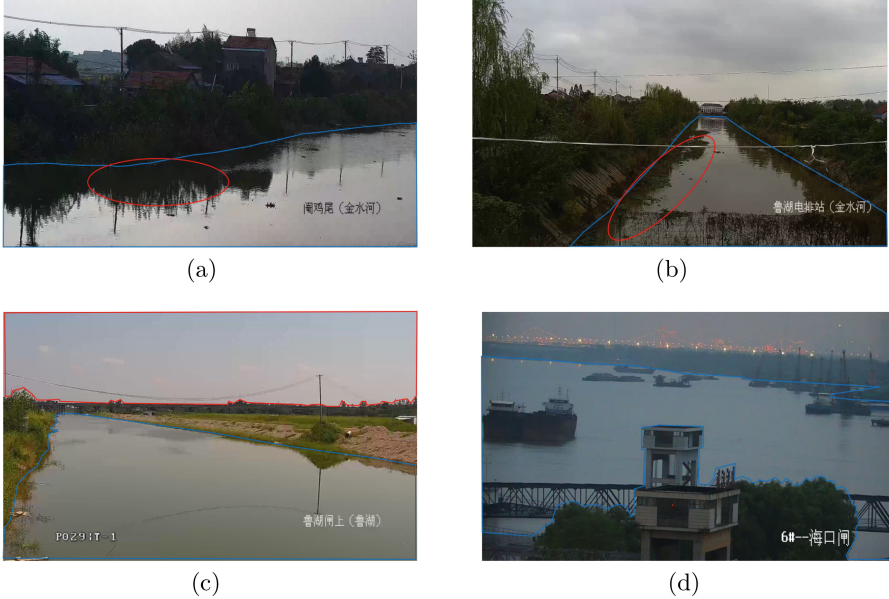
Inland water is one of the most protected resources. However, Water pollution threatens the health of the water quality, and will cause a large number of dead fish, cyanobacteria outbreaks and other ecological disasters. At the same time, floods and dry waters will also cause a huge threat to human society. Therefore, it is important to supervise the water surface situation. To monitor the water environment, identifying the water region and segment it from its surroundings is essential.

---

This work was supported by the Science and technology project innovation fund of Hubei Three Gorges Laboratory under Grant SC215002, National Natural Science Foundation of China under Grant 62072350, Grant 62171328; and the Hubei Technology Innovation Project under Grant 2019AAA045.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
G. Zhai et al. (Eds.): IFTC 2022, CCIS 1766, pp. 217–228, 2023.  
[https://doi.org/10.1007/978-981-99-0856-1\\_16](https://doi.org/10.1007/978-981-99-0856-1_16)

As shown in Fig. 1, due to the specularity and indeterminacy of water, problems such as reflection on the water surface can, weather changes, and lighting changes will significantly change the feature distribution on the water surface. Hence, we thought of use semantic segmentation for water segmentation.



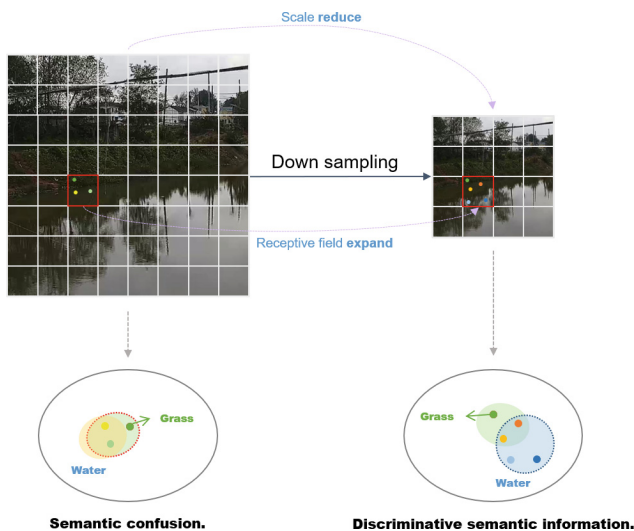
**Fig. 1.** Some common situation in water segmentation. The first row is two images of the water surface on a cloudy day. In (a), the severe reflection blurs the boundary between the water surface and the plants. In (b), Floating objects such as algae, garbage, and reflections coexist in the water, complicating the situation on the water surface. (c) and (d) show the conditions of the water surface on sunny days and in the toward evening, respectively, which make the distribution of water surface features more uncertain.

Semantic segmentation or pixel-level classification [1], which aims to assign each pixel of an image to one category, is one of the key problems in computer vision. Semantic segmentation is applied in many scenes such as geographic information systems and medical image analysis systems, advanced driver assistance systems (ADAS) and various applications in autonomous vehicles. Most of the semantic segmentation methods reproduce the details in a fine way [16, 21], so that each pixel can be segmented under the condition of fuzziness and occlusion. However, in the task of water segmentation, many interference details (such as various reflections on the water surface and the appearance of various underwater plants) may lead to network learning irrelevant semantic to the nature of water, thus resulting in inaccurate segmentation. In order to solve the above problem, a natural idea is to learn features of a large area with the help of a larger scale receptive field and reduce the interference of local harmful information Fig. 2.

In this paper, we analyze the characteristics of the task of water segmentation in the scene of inland rivers in Hubei Province under the monitoring video. We propose a asymmetric interaction module (AIM) to converge the water features to a larger receptive field and use this powerful module to construct asymmetric multiscale interaction network (AMINet). The proposed network uses AIM to fuse the features of various scales, so as to achieve the goal of accurate learning of large scale features to segment the water boundary more accurate.

Our contributions are mainly in the following folds:

- (1) We propose a new semantic segmentation dataset, Hubei Water dataset, which includes water monitoring images of different time periods and weather in Hubei and annotations of the corresponding water surface and surrounding environment.
- (2) We propose a novel multiscale semantic segmentation network, unidirectional feature finege network (AMINet), which converge features into large scale features with our proposed asymmetric interaction module (AIM). Within the scope of our knowledge, this is the first work on semantic segmentation of water surface in surveillence scenarios.
- (3) We empirically demonstrate the superior key point detection performance the Hubei water dataset we collected. Compared with existing popular CNN-based semantic segmentation methods, we achieve a 5.9-point improvement. In addition, we have achieved unimaginable improvements in visual effects.



**Fig. 2.** With the same size red patch, the point in left image may confuse whether it belongs to water or grass. But in right one, grass point and water point can be will separate. Enriching images with large receptive fields is beneficial to water segmentation. (Color figure online)

## 2 Related Work

**Semantic Segmentation.** Due to the rapid development of deep learning [2–10], Fully Convolutional Networks (FCNs) [11] have been a dominated and fundamental work in the field of semantic segmentation. However, only convolution-based architectures are difficult to handle large objects due to the weak ability of convolution operations to process global information. To alleviate this problem and enhance the global correlation capability of convolution-based architectures, LC Chen et al. propose atrous convolution [12, 13], Zhao et al. propose a pyramid pooling module [14]. Meanwhile, follow the [15], Huang et al. find that attention mechanism can effectively change the network preferences [16]. After that, Yuan et al. use the area contextual information to solve context aggregation problem [21].

**Water Segmentation.** Water segmentation plays a supporting role in water area monitoring and water quality warning. To achieve water segmentation of monitoring scenes, traditional methods mostly rely on low-level features. By using decision forests [17] or support vector machines [18] on low-level features, one can achieve simple water segmentation task. To improve the segmentation accuracy, Kristan proposed a method [19] to use inertial measurement unit to assist maximize expectations for water segmentation. Further more, Lopez-Fuentes et al. proposed a simple CNN method [20] to detect flooding in rivers by water segmentation.

The existing research on water segmentation in monitoring scenes is not yet mature. Our method aims to achieve high-precision water semantic segmentation in monitoring scenes. We choose the representative CCNet [16] and the strong semantic relevance network OCRNet [21] as the comparison objects.

## 3 Methodology

In this section, we first display the proposed asymmetric multiscale interaction Network (AMINet). Then, We introduce the core component of AMINet, asymmetric interaction module (AIM), shown in Fig. 3. After that, we introduce the loss we use. At last, we introduce the proposed water segmentation dataset, Hubei water dataset.

### 3.1 Asymmetric Multiscale Interaction Network

We focus on the design of the main body and introduce our asymmetric multiscale interaction network. The goal of this network is, given an input image with size of  $H \times W \times 3$ , we generate different scale feature map set  $F_i$  with a resolution of  $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ , where  $i \in 1, 2, 3, 4$ . Then, By interacting information between layers, a mask with both semantic information and detailed information is generated. **Sequential Feature Enrichment Subnetworks.** Existing CNN semantic segmentation networks are constructed by concatenating sub-networks of different resolutions, where each sub-network forms a stage, which

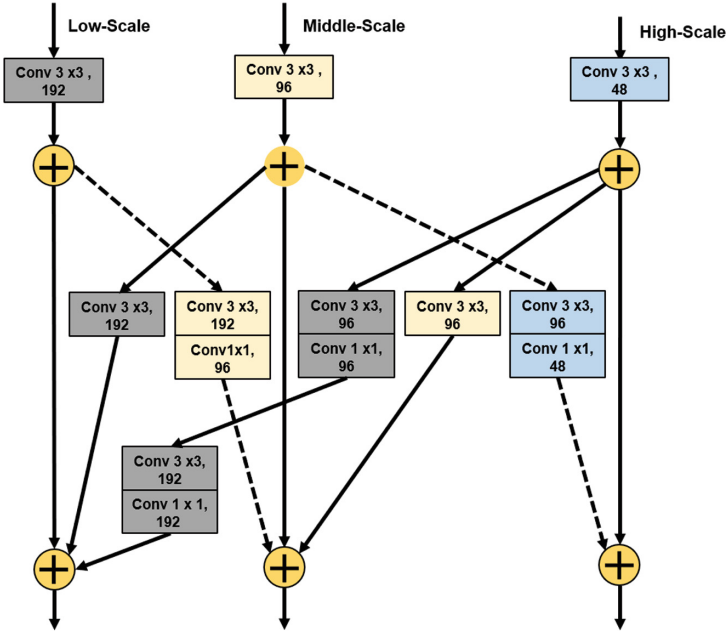
consists of a series of convolutions, and adjacent sub-networks are up-sampling or down-sampling to separate the resolution. The rate increases or decreases in multiples. In the next paragraphs of this section, we detailed the architecture of our proposed wide area enhanced multi-scale feature fusion network and the asymmetric interaction module.

**Table 1.** The architectures of AMINet.

	Output size	Asymmetric multiscale interaction network			
Stage1	$1024 \times 1024$	$3 \times 3$ 32			
Stage2		$3 \times 3$ 32		$3 \times 3/2$ 64	
	$1024 \times 1024$	$3 \times 3$ 32		$3 \times 3$ 64	
	$512 \times 512$	$3 \times 3$ 32		$3 \times 3$ 64	x 2
		Low to high fusion			
Stage3	$1024 \times 1024$	$3 \times 3$ 32		$3 \times 3/2$ 64	$2 \times [3 \times 3/2]$ 128
	$512 \times 512$	$3 \times 3$ 32		$3 \times 3/2$ 64	$2 \times [3 \times 3/2]$ 128
	$256 \times 256$	$3 \times 3$ 32		$3 \times 3$ 64	$3 \times 3/2$ 128
Stage4			$3 \times 3/2$ 64	$2 \times [3 \times 3/2]$ 128	$3 \times [3 \times 3/2]$ 256
		$3 \times 3$ 32	$3 \times 3$ 64	$3 \times 3/2$ 128	$2 \times [3 \times 3/2]$ 256
					$3 \times 3/2$ 256
	$1024 \times 1024$	$5 \times 5$ 32	$5 \times 5$ 64	$5 \times 5$ 128	$5 \times 5$ 256
	$512 \times 512$	Low to High fusion			
	$256 \times 256$			$2 \times [3 \times 3/2]$ 128	$3 \times [3 \times 3/2]$ 256
	$128 \times 128$	$5 \times 5$ 32	$3 \times 3/2$ 64	$3 \times 3/2$ 128	$2 \times [3 \times 3/2]$ 256
		$5 \times 5$ 64	$5 \times 5$ 128	$3 \times 3/2$ 256	
				$5 \times 5$ 256	
Output	$1024 \times 1024$	Low to high fusion			

**Network Construction.** With a large-scale layer as the first layer, our network propagates construct layer by layer while maintaining high-resolution features. As a result, the layers in the later stage consists of resolution from the previous stage and an extra lower resolution. This process can be vividly described as a inverted triangle construction process. As the Table 1 shown, our network consist of four stages. They can be subjectively divided into two part, stage 1 to stage 3 can be considered as a whole to obtain effective features of each scale, and stage 4 is to integrate features of each resolution into low resolutions to obtain a wide-area enhancement scale feature. In this table,  $K \times K/S$  represents a layer with a convolution kernel size of  $K$ , stride size of  $S$  and its supporting BN Layer and ReLU Layer. To be specific, given an image of size  $H \times W \times 3$ , we first resize the image scale into size  $1024 \times 1024$ . With a multi-resolution feature extraction network, the information of the image gradually convergence on low-resolution feature maps, i.e. feature maps with a wider receptive field.

### 3.2 Asymmetric Interaction Module



**Fig. 3.** The architecture of asymmetric interaction module. The ReLU and BN layers after every Conv are hidden.

An example of the detailed architectures of asymmetric interaction module is shown in Fig. 3. For a three-branched input, a series bottleneck is added at the end of each branch. The cross fusion includes fusing the high-scale branch into low-scale branch (high-to-low fusion) and fusing the adjacent low-scale branch into high-scale fusion (low-to-high fusion). For low-to-high fusion, low-scale feature maps are first compressed by a  $3 \times 3$  convolution and then upsampled by a  $1 \times 1$  convolution. For high-to-low fusion, high-scale feature maps are downsampled by a series of  $3 \times 3$  convolution with a stride of 2 and  $1 \times 1$  convolution. For the  $i$ -th size  $N \times N$  feature map, the fusion feature map  $X_{i,N}$  can be written as:

$$X_{i,N} = \sum FH(X_{i-1,N \times j}) + FL(X_{i-1,N/2}) + ConvBlock(X_{i-1}, N) \quad (1)$$

where  $FH$  and  $FL$  refer to the feature from high scale feature maps and low feature maps, the  $ConvBlock$  represent a sequence of convolutional layer with BN and RELU.

### 3.3 Post Processing

In this part, we use the strategy used in OCRNet to make our. Unlike other secondary processing algorithm such as ASPP, etc. which sample around the

target point, OCRNet use the object segmentation area to replace the sparse point. In this paper, This strategy is added after the AMINet to better focus on area context information.

### 3.4 Loss

In this paper, we only adopt simple extra supervision for a fair comparison with most of the methods. Following the PSPNet, we add the auxiliary loss and set the weight to 0.4. The final loss can be expressed as:

$$L_f = L_c + \alpha L_a u x \quad (2)$$

where  $L_f$  and  $L_c$  are the final loss and the cross-entropy loss,  $L_a u x$  represents auxiliary loss with a weight  $\alpha = 0.4$ .

### 3.5 Hubei Water Dataset

**Construction.** The main types of water surface in the study area are rivers and lakes. To construct this dataset, we collected data from surveillance videos of different waters in Hubei Province. In order to obtain samples in different weather and at different times, we collected samples in three time periods in different weather, namely, 9:00 a.m.–11:00 a.m., 11:30 p.m.–1:30 p.m. and 6:00 p.m.–8:00 p.m.

**Dataset Scale and Partition.** Hubei water dataset consists of 896 images with a size of  $2560 \times 1440$ . images are evenly divided into training (598), verification (151) and test sets (147) according to the scene.

**Pre-processing.** The images with low imaging quality and the images after 7:30 pm cannot identify any effective information through human eyes or the network, so they are removed in this dataset.





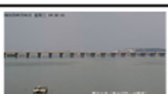






**Class Selection Rules.** In order to maximize the difference between the reflection on the water surface and the entities out of the water, we selected other nine classes to help segment the water region. The Fig. 4 display the classes we select and the reason for choosing.

**Dataset Label Generation.** Labels are interpreted as a polygon shape file format showing the water area. We carefully separated the water surface boundary, summarized the common reflections and floating objects on the water surface, and then marked them manually.

## 4 Experiments

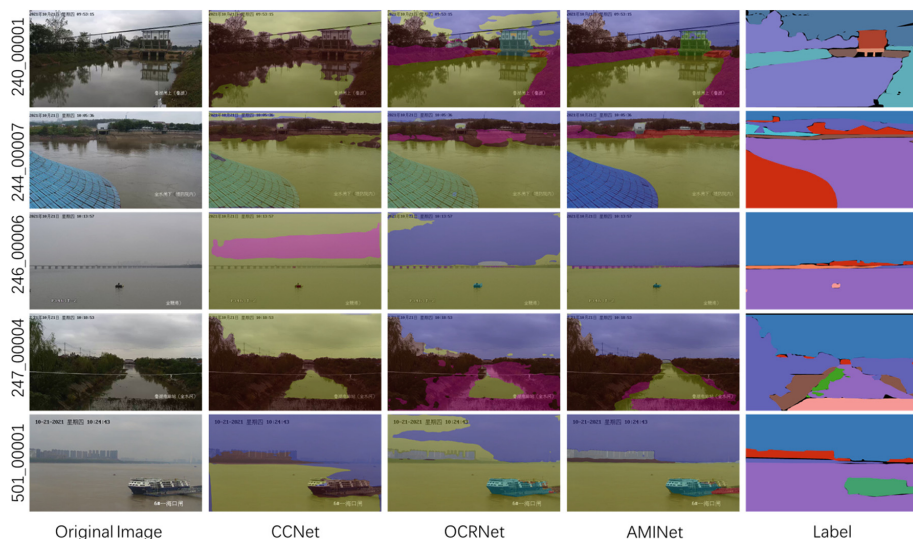
### 4.1 Dataset and Metrics

**Dataset.** In this part, We train and evaluate our model on the aforementioned Hubei water dataset.

id	class	discription	example
0	water	Refers to the inland waterways of the Yangtze River basin, including lakes and rivers. The target class	
Major classes			
1	algae	A common floating object on the water surface. It has become a major pest of rivers and dams.	
2	sky	The sky above the water. The reflection of sky is always exists on water surface	
3	tree	Waterfront trees. The most difficult reflection is usually from a tree.	
Other classes			
4	bridge	A structure built to span the body of water. It always been the boundary of water and sky	
5	boat	Transportations on waterways such as rivers and lakes. Often appears in the dataset	
6	building	Waterfront buildings (such as houses). Often appears in the dataset	
7	shore	The fringe of land at the edge of a large body of water. Cause server reflection	
8	sluice	A movable gate allowing water to flow under it. Often appears in the dataset	
9	road	It is found in training that this kind of water is easily confused with water surface.	
10	grass	It is found in training that this kind of water is easily confused with water surface	

**Fig. 4.** Total classes of the Hubei water dataset. The class names and the choose reason are given. Examples are given in the end of each row.





**Fig. 5. Qualitative results on Hubei water dataset.** compared to CCNet and OCRNet, our AMINet predicts masks with substantially finer details near water boundaries. And our network can better distinguish the reflection on the water surface.

**Evaluation Metrics.** Although the dataset labeled about eighteen categories, but the goal is to separate the water region. So we take the Intersection over Union (IOU) as the final inspection index.

## 4.2 Implementation Details

**Training Sets.** We trained with the input size of  $1024 \times 1024$ . The data augmentation includes random crop and random rotation. Only water is taken as positive sample. We use stochastic gradient descent (SGD). The base learning rate is set as  $1e - 2$ , and is dropped to  $1e - 4$  and  $1e - 5$  at the 130th and 176th epoch. The training processing is terminated within 200 epochs.

**Test Sets.** We binarize the labeled mask images of pixel classification results and test sets by whether they are water, and then calculate water IOUs.

## 4.3 Result on Test Set

We report the results of our method with other advanced methods. AS shown in Table 2, Our method get a 78.5% IOU and which surpasses other advanced method. Figure 5 shows qualitative results on Hubei water dataset, where AMINet provides better effect and details than CCNet and OCRNet.

**Table 2.** Comparison on the Hubei water test set. The best result is in bold.

Method	Backbone	Input size	IOU
CCNet	ResNet-101	$2560 \times 1440$	43.2
HRNet	HRNet-W48	$1024 \times 1024$	68.3
OCRNet	HRNet-W48	$1024 \times 1024$	72.6
AMINet	AMINet	$1024 \times 1024$	<b>78.5</b>

#### 4.4 Ablation Study

In this part, we separate our network into three part: The AMINet (/) use a simple bilinear to replace low-to-high fusion and use only convolutional layer with BN+RELU to replace high-to-low; The core component asymmetric interaction module; The Conv $5 \times 5$ . We analyze the effect of these factors and the direct effect is shown in Table 3.

**Effects of AIM.** from the Table 3, it is easy to find that AIM greatly improves the IOU, which proves the effectiveness of this module and focuses more attention on the characteristics of larger receptive field to water segment.

**Conv  $5 \times 5$ .** After using  $5 \times 5$  convolution in the last several layers, the IOU is 2.4 % and 1.8 % higher than that of AMINet (/) and AMINet (/)+AIM respectively. This proves that large convolution kernels will have better effect for large water areas with uneven characteristic distribution.

**Table 3.** The effects of factors in AMINet.

AMINet (/)	+AIM	+CONV $5 \times 5$	IOU $\uparrow$
✓			71.4
✓	✓		76.7
✓		✓	73.8
✓	✓	✓	<b>78.5</b>

## 5 Conclusion

In this paper, we propose AMINet, a powerful water semantic segmentation method and introduce our dataset Hubei water dataset. With the effective asymmetric interaction module, our framework gradually aggregates information on the smallest feature layer with a larger perceptive field, and ultimately gets better performance on the Hubei water dataset. The disadvantage is that AMINet does not achieving good results in all categories. We leave it in the future.

## References

1. Li, Z., Wang, R., Zhang, W., Hu, F., Meng, L.: Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. *IEEE Access* **7**, 155787–155804 (2019)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
4. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519 (2019)
5. Wang, W., Li, X., Yang, J., Lu, T.: Mixed link networks. *arXiv preprint arXiv:1802.01808* (2018)
6. Lu, T., Wang, Y., Zhang, Y., Jiang, J., Wang, Z., Xiong, Z.: Rethinking prior-guided face super-resolution: a new paradigm with facial component prior. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
7. Wang, Y., Lu, T., Zhang, Y., Wang, Z., Jiang, J., Xiong, Z.: FaceFormer: aggregating global and local representation for face hallucination. *IEEE Trans. Circuits Syst. Video Technol.* (2022). <https://doi.org/10.1109/TCSVT.2022.3224940>
8. Lu, T., et al.: Face hallucination via split-attention in split-attention network. In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5501–5509 (2021)
9. Wang, Y., Lu, T., Zhang, Y., Fang, W., Wu, Y., Wang, Z.: Cross-task feature alignment for seeing pedestrians in the dark. *Neurocomputing* **462**, 282–293 (2021)
10. Wang, Y., Lu, T., Zhang, Y., Wu, Y.: Multi-scale self-calibrated network for image light source transfer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 252–259 (2021)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015)
12. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2017)
13. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017)
14. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890 (2017)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
16. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: criss-cross attention for semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 603–612 (2019)
17. Yao, T., Xiang, Z., Liu, J., Xu, D.: Multi-feature fusion based outdoor water hazards detection. In: *2007 International Conference on Mechatronics and Automation*, pp. 652–656. *IEEE* (2007)

18. Achar, S., Sankaran, B., Nuske, S., Scherer, S., Singh, S.: Self-supervised segmentation of river scenes. In: 2011 IEEE International Conference on Robotics and Automation, pp. 6227–6232. IEEE (2011)
19. Kristan, M., Kenk, V.S., Kovačič, S., Perš, J.: Fast image-based obstacle detection from unmanned surface vehicles. *IEEE Trans. Cybern.* **46**(3), 641–654 (2015)
20. Lopez-Fuentes, L., Rossi, C., Skinnemoen, H.: River segmentation for flood monitoring. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 3746–3749. IEEE (2017)
21. Yuan, Y., Chen, X., Chen, X., Wang, J.: Segmentation transformer: object-contextual representations for semantic segmentation. arXiv preprint [arXiv:1909.11065](https://arxiv.org/abs/1909.11065) (2019)